面向食品安全突发事件汉语分词的特征选择及模型优化研究*

张　越[1]　王东波[1,2]　朱丹浩[3]

[1](南京农业大学信息科学技术学院　南京 210095)

[2](南京农业大学领域知识关联研究中心　南京 210095)

[3](江苏警官学院图书馆　南京 210031)

摘要: 【目的】在食品安全领域中, 建立相关数据库对食品安全的监管和控制都会有很大的帮助, 自动分词在构建索引、使用索引以及构建语料库中都起到至关重要的作用。将基于条件随机场的字标注统计学习方法, 应用在食品安全突发事件语料的自动分词中。【方法】分析语料的词长分布等特点, 对该方法自动分词过程中所涉及的特征选择和特征模板进行不同实验, 得出不同特征选择和应用不同特征模板对分词结果的影响。【结果】从实验结果可以看出, 特征选择时并不是特征越多分词效果越好, 会出现特征干扰的情况, 在二三字词占 46.62%的食品安全突发事件语料中, 特征模板中的当前字和前后驱第一个字所代表的特征模板对分词效果影响明显。【结论】通过对不同特征选择和特征模板及其相互组合的实验, 选择出在本文研究的语料库自动分词中最优的特征和特征模板, 在 5Tag 特征标记下配合对应特征模板对目标语料分词的 F 值达到 92.88%。

关键词: 中文分词　食品安全　条件随机场　特征模板　特征选择

分类号: G351

1　引　言

近年来, 食品安全事故不断涌现。越来越多的食品安全恶性事件, 对社会生产和人民群众的生活造成严重影响。关于食品安全突发事件的各种信息也迅速增多, 并引起人们的广泛关注, 由于食品安全关系到民众的生命安全和公共健康, 因此, 食品安全问题的解决不仅需要“自上而下”政府部门的行政监管和企业的自律[1], 更需要“自下而上”社会监督力量的积极参与[2]。在信息传播速度如此之快的今天, 作为一直以来的社会热点话题, 网络、纸质报纸、书籍成为“食品安全突发事件”快速扩散的主要载体, 同时也成为群众获取食品安全事件信息的一个主要途径。随着自然语言处理的蓬勃发展, 针对中文文本的自动分词技术的研究已取得一定的成效, 在精准度和分词速度上都有了大幅提升, 这项技术在许多方面也已经得到应用, 在文本分类、信息检索、信息过滤、文献自动标引、摘要自动生成等中文信息处理中都起到关键性的作用[3], 但是在食品安全信息处理中自动分词的应用和研究较少, 有待探索。

在食品领域中, 建立相关数据库对食品安全的监管和控制会有很大的帮助, 张星联等[4]指出建立食品安全预警数据库系统的重要性。食品信息的不对称不真实是食品领域中的不正当行为的根本原因之一, 要避免这样的现象, 就要创立有效可行的食品电子监管系统, 也就是创建动态数据库, 数据库更新及时和准

通讯作者: 王东波, ORCID: 0000-0002-9894-9550, E-mail: db.wang@njau.edu.cn。

*本文系国家自然科学基金项目“基于 CSSCI 的句法级汉英平行语料库构建及知识挖掘研究”(项目编号: 71303120)、2011 协同中心项目“面向应急推演平台的海量突发事件知识库与模型库构建研究”(项目编号: JD20150101)和江苏省高校哲学社会科学项目“高校危机管理案例知识库构建及知识挖掘研究”(项目编号: 2014SJB246)的研究成果之一。

确以及透明可以确保行业的食品更加安全，秩序也更加稳定[5]。食品工业发展迅速，食品加工的范围和深度不断扩展，余清等[6]分析了建立加工食品风险数据库的必要性，该数据库可以提供食品的检验和检测信息，还提供如食品的危害物风险系数等信息，为加工食品风险数据的研究提供了很大的帮助。在具体实施上，贾凯等[7]建立了彭州市三界镇生鲜农产品溯源数据库，在对国内和国外的溯源系统进行整理和研究的基础上，再对彭州市三界镇的具体食品情况进行整理，并提供生鲜食品信息的管理和应用功能。

目前中文自动分词方法主要有 4 种：机械分词法、基于统计的分词法、字标注统计学习法以及基于深度神经网络模型的方法。在 2002 年之前，自动分词方法基本上是基于词典的[8]，在此基础上可进一步分为基于规则的机械分词法和基于统计的分词法。这两类方法完全依赖于词典，词典内容则是全部领域信息的来源[9]，虽然该方法配合词典以及通过补充大量消除歧义的信息，能够有较好的领域针对性和准确率，但是其受限于对词典的完全依赖，导致这两类方法不能够有很好的适应性，另外构建领域词典工程量大，大量的时间和人力花在词典构建上，同时随着更多未登录词的出现，词典难以维护。

随着 SIGHAN 国际中文分词评测 Bakeoff 的展开，将中文分词任务视为序列标注问题逐渐成为主流。字标注统计学习方法在解决未登录词和消除歧义上有较好的效果，在不利用词典的情况下，字标注统计学习方法的分词效果完全超过基于词典的方法，显然是更好的选择。而基于深度神经网络模型的方法，目前尚未成熟，深度学习在自然语言处理方面的应用较少，本文不做探讨。

基于字标注统计学习方法的中文分词任务本质上是一个序列标记的过程，将文本信息抽象为一个观察序列，然后对序列中的每个字进行标记[10]。字标注统计学习方法的关键在于选择一个对处理目标合适的机器学习模型，而目前用的比较多的是隐马尔可夫模型(HMM)、最大熵模型(ME)和条件随机场模型(CRF)。隐马尔可夫模型主要缺点是由于其输出独立性假设，导致不能考虑上下文的特征，限制了特征的选择。最大熵模型则解决了隐马尔可夫模型的问题，可以任意选择特征，但由于其在每一节点都要进行归一化，所以只能找到局部的最优值，但是也带来了标记偏见的问题，即凡是训练语料中未出现的情况全都忽略掉。条件随机场模型则很好地解决了这一问题，该模型并不在每一个节点进行归一化，而是所有特征进行全局归一化，因此求解的是全局最优值。在之前的研究已经证明采用链式的 CRF 模型实现的分词系统，较之于 ME 与 HMM 能得到更好的效果。

本文根据食品安全突发事件语料特点，提出一种面向食品安全突发事件汉语分词的特征选择及模型优化的研究方法。研究内容侧重于以下两个方面：将基于链式的条件随机场模型的中文自动分词方法应用于食品安全语料自动分词当中；分析语料，提出符合语料特点的特征模板、特征选择以及特征标记选择。该方法与其他分词系统相比能够较好地解决食品安全案例库这种密集型文本所具有的交叠歧义和未登录词的问题，有效提高了分词的准确率和召回率。

2　条件随机场模型介绍

条件随机场模型(Conditional Random Fields, CRFs)是 Lafferty 等于 2001 年在最大熵模型和隐马尔可夫模型的基础上提出的一种无向图学习模型，是一种用于标注和切分有序数据的条件概率模型[11]。

无向图模型亦称为马尔可夫随机场或马尔可夫网络，是由 Pearl 提出[12]。无向图 $G(V,E)$，其中 V 是顶点/节点，表示随机变量；E 是边/弧，表示随机变量间的条件依赖关系。

尽管在给定每个节点的条件下，分配给该节点一个条件概率是可能的，无向图的无向性导致不能用条件概率参数化表示联合概率，而要从一组条件独立的原则中找出一系列局部函数的乘积来表示联合概率。

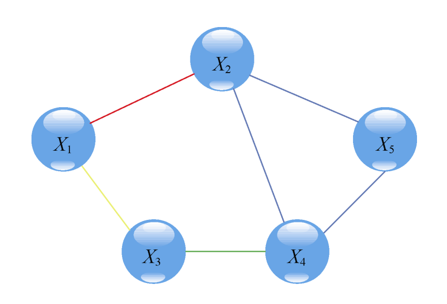


图 1　无向图最大全联通子图示例

图 1 是一个简单的例子，无向图中的最大全联通

子图为$\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_3, X_4\}$, $\{X_2, X_4, X_5\}$，那么很容易得到图中无向图模型的联合概率分布为：

$$P(X_1,X_2,X_3,X_4,X_5)=\frac{\Psi_1(X_1,X_2)\Psi_2(X_1,X_3)\Psi_3(X_3,X_4)\Psi_4(X_2,X_4,X_5)}{\sum_{X_1,X_2,X_3,X_4,X_5}\{\Psi_1(X_1,X_2)\Psi_2(X_1,X_3)\Psi_3(X_3,X_4)\Psi_4(X_2,X_4,X_5)\}} \quad (1)$$

如果给定的马尔可夫随机场中每个随机变量还有观察值，则要确定的是给定观察集合下，这个马尔可夫随机场的分布，也就是条件分布，这个马尔可夫随机场就称为条件随机场。它的条件分布形式完全类似于马尔可夫随机场的分布形式，只不过多了一个观察集合 X。

条件随机场提出目的在于解决离散数据的序列标注问题，在给定的序列 $X=\{x_1, x_2, x_3\cdots, x_{n-1}, x_n\}$和有限状态集合 $Y=\{y_1, y_2, y_3\cdots y_{n-1}, y_n\}$的情况下，设 $G=(V, E)$是一个无向图，$Y=\{Y_v|v\in V\}$是以 G 中节点为索引的随机变量构成的集合。在给定 X 的条件下，如果每个随机变量服从马尔可夫属性即：

$$P(Y_v|X,Y_u,u\neq v)=P(Y_v|X,Y_u,u\backsim v) \quad (2)$$

其中，$u\backsim v$ 表示 u 和 v 是相邻的边，则构成一个条件随机场。

如图 2 所示，CRF 采用无向图模型来描述给定序列的状态，在条件随机场 X 中，每一个元素对应图中的每一个节点，而每一条边则代表每一个节点的状态，CRF 就是给定观察集合情况下的无向图模型[13]。

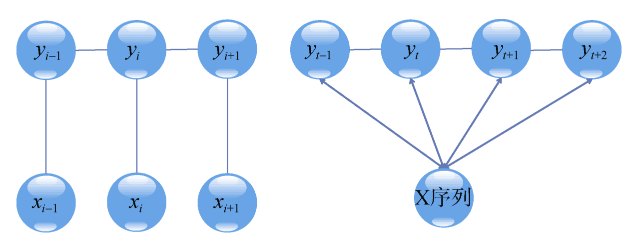


图 2　线性链的 CRF 图形结构

根据条件随机场基本理论：

$$P(y|x,\lambda)\propto \exp(\sum_j \lambda_j t_j(y_i-1,y_i,x,i)+\sum_k u_k s_k(y_i,x,i)) \quad (3)$$

3　基于 CRF 模型的食品安全突发事件自动分词

3.1　食品安全语料库说明

在对食品安全突发事件进行采集、标注和组织的基础上，本文构建 2005 年–2015 年的食品安全突发事件语料库，并由此语料库经过粗切分、人工校对等步骤形成最后实验数据，具体过程如图 3 所示。

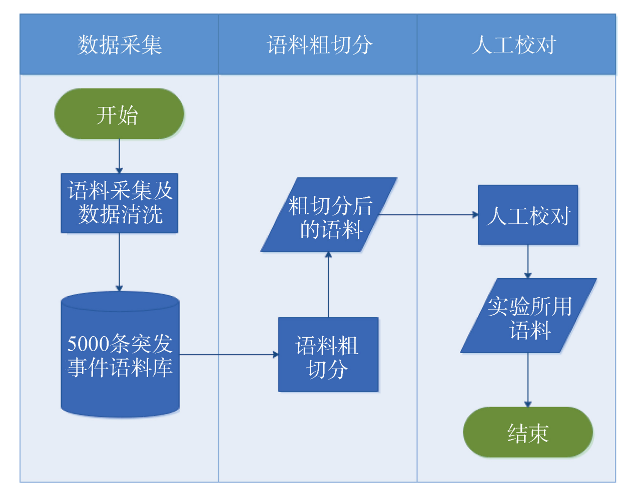


图 3　实验所用语料库构建过程

(1) 食品安全突发事件的采集过程如下：采集目标主要包括网络上的食品安全突发事件和纸质报纸、书籍上的食品安全突发事件。网络上食品安全突发事件的采集通过自己编写的程序，利用面向突发事件主题垂直搜索引擎技术自动采集，采集范围包括新闻门户、论坛和博客，对于采集的异构数据通过相应的数据清洗、转换保存到数据库中；而纸质的突发事件案例则通过人工录入、校对的方式完成对近 5 000 条突发事件的采集，经清洗和转换后约 4 500 条入库数据，累计存储大小约 20MB。

(2) 采集完成之后，用中国科学院计算技术研究所的分词软件 NLPIR 对语料进行标注，标注结果显示，其结果中出现了很多未登录词识别不准确的情况。食品安全案例库属于密集型文本，其中中文未登录词和歧义词大量出现。在为数众多的食品描述、地理描述、化合物描述等类型文本中，食品安全描述文本具有很好的代表性，涉及众多的食品名称和化学品名称，所以机器会标注错误，在食品安全领域适用性不太好，所以只是用它来进行粗切分，减少人工标注的工作量。

(3) 对经过粗切分之后的语料进行人工标注，由于出现了较多的未登录词未识别和歧义词识别错误的情况，因此对全部粗切分后的语料逐个词进行校对，找出粗切分过程中出现的分词错误，并校正为正确的

分词结果，最大程度上保证训练语料分词的正确。

3.2 实现方法

CRF++[14]是一个可用于连续序列的标注的可定制并且开源的条件随机场工具，而且也是目前所有条件随机场工具中使用率最高，被普遍认为易用性、准确性和稳定性等综合方面表现最好的一个。CRF++是为了通用目的设计定制，并被用于自然语言信息处理的各个方面，如命名实体识别、信息提取、语义分析等。本文利用 CRF++进行中文文本分词处理，使用的版本是 CRF++在 Linux 环境下较新的 0.58 版本。

实验过程如图 4 所示，主要分为训练学习、测试输出、模型测评和模型优化 4 个阶段。训练学习部分主要是语料的特征提取，选出适合食品安全语料的部分特征，将不同特征赋予不同特征标记之后加入文本中并处理成 CRF++能识别的格式。然后选取不同特征及不同特征组合，根据选取的特征构造特征模板。最后在 CRF++中由训练数据和特征模板(template 文件)一起训练出分词模型(model 文件)。测试输出部分是用同样是处理为 CRF++格式的测试数据和已经训练出的分词模型共同得到最后的分词结果，如表 1 所示。模型测评部分对得到的输出结果进行测评，并将测评结果与其他实验结果进行对比，观察其中差别，不断改变特征选择、特征标记的选择以及优化特征模板，直到得到相对最优的分词结果。

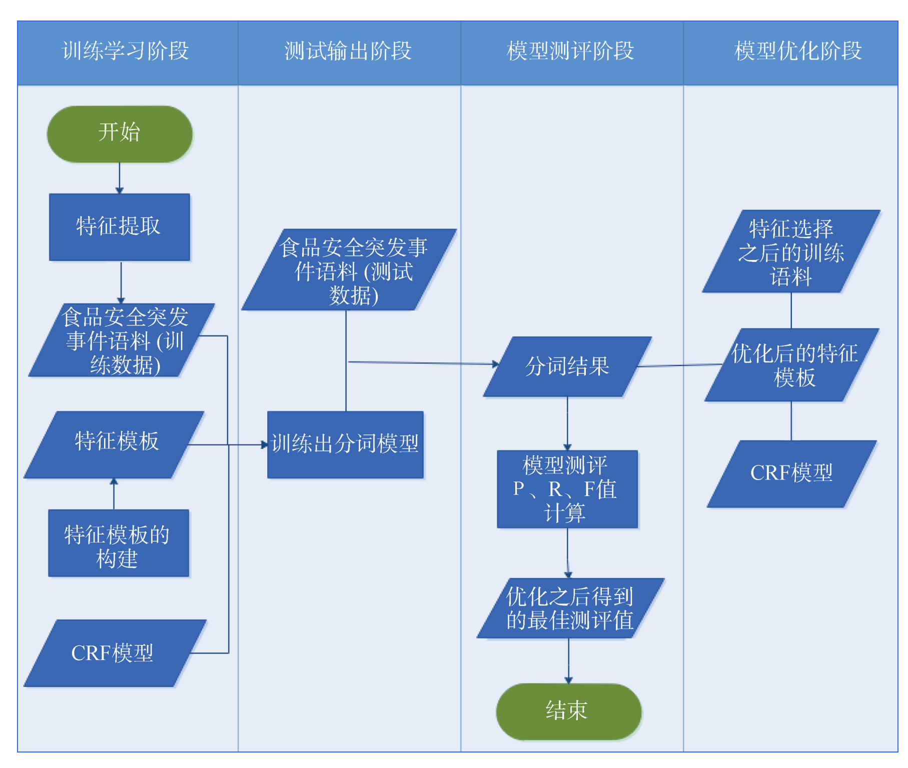


图 4 实验流程

表 1 CRF 分词后输出结果示例

文本语料	正确标记	CRF 输出标记	文本语料	正确标记	CRF 输出标记
另	S	S	中	S	S
一	B	S	麻	B	B
种	E	S	痹	I	E
是	S	S	大	M	B
生	B	B	意	E	E
产	E	E	。	S	S

由于本文的语料库较为庞大，而且原始语料无任何标注，因此需要对所有参与训练和测试的语料进行分词。采用的方法是机器自动分词与人工校正相结合的方法，首先用汉语分词系统 NLPIR 进行自动分词[15]，由于食品安全事件语料的领域性较强，对于 NLPIR 错误的分词结果，在相应食品科学领域研究人员的指导下，组织人力系统全面地对食品安全事件的语料进行分词校对，形成高质量的食品安全事件分词语料[16]。本文所有的实验均基于人工校对后的分词语料进行。

食品安全突发事件语料的中文自动分词可以抽象为序列标记任务，输入的待分词文本就是模型中给定的观测序列，在给定观测序列的条件下，利用 CRF 模型得到一个整个序列的最大的联合概率分布。能否选出一组有效的特征标记对最后的分词效果有很大影响，所以首先要筛选和确定待分词语料的特征选择和相对应的特征标记，然后根据筛选的特征确定训练模型时所用到的特征模板，训练数据和特征模板确定之后，就可以训练出所需要的 CRF 模型，最终的分词结果即是根据该模型计算得到的。

模型优化阶段是本文的重要研究部分，通过不断尝试新的特征标记以及不同的特征选择组合，配合以与文本特点和特征选择更加适合的特征模板以达到更好的模型测评结果，具体表现在 P(准确率)、R(召回率)、F 值更高。具体计算公式如下：

$$准确率(P)=\frac{A}{A+B}\times 100\% \quad (4)$$

$$召回率(R)=\frac{A}{A+C}\times 100\% \quad (5)$$

$$调和平均值(F)=\frac{2\times P\times R}{P+R}\times 100\% \quad (6)$$

实际计算的时候也是如此，对位置特征的标记计算 P、R、F 的值，然后，根据每个标记数量所占比例计算出权值，加权平均得到最后的值。

(1) 特征和特征标记的选择

在基于条件随机场的中文自动分词中，在训练学习阶段，需要给定一部分正确的经过机器分词和人工校正的语料来训练学习出用于分词的 CRF 模型，在训练语料中使用不同的特征选择和不同的特征标记会导致分词效果的不同。

本文使用现代汉语中字在词中的位置特征，测试了三种不同的位置特征标记，如表 2 所示，同时将位置特征放在训练语料的最后一列，作为 CRF++的输出，用测试语料配合经过训练的模型分词之后输出位置特征，最后根据位置特征标记将字组成词，完成分词任务。

表 2　位置特征标记

标记类型	标记描述
4Tag {B, M, E, S}	B 表示词首字, M 表示词中字, E 表示词尾字, S 表示单字词字。
5Tag {B, I, M, E, S}	B 表示词首字, I 表示四字以上词首后第一个字, M 表示词中, E 表示词尾字, S 表示单字词字。
6Tag {B, I, J, M, E, S}	B 表示词首字, I 表示四字以上词首后第一个字, J 表示五字以上词首后第二个字, M 表示词中字, E 表示词尾字, S 表示单字词字。

在计算 P、R、F 值的时候需要给每个标记的一个权值，本文对其权值的计算，如表 3–表 5 所示，统计出每个特征标记在测试语料中的数量，计算其所占比例，用该标记数量在总标记中的百分比作为其权值。

表 3　4Tag 特征标记数量情况

特征标记	标记数量	标记所占百分比
B	597 343.7	30.22%
M	158 744.3	8.03%
E	597 343.7	30.21%
S	623 538.5	31.54%

表 4　5Tag 特征标记数量情况

特征标记	标记数量	标记所占百分比
B	597 343.7	30.22%
I	28 529	1.44%
M	130 215.3	6.59%
E	597 343.7	30.21%
S	623 538.5	31.54%

表 5　6Tag 特征标记数量情况

特征标记	标记数量	标记所占百分比
B	597 343.7	30.22%
I	28 529	1.44%
J	11 595.4	0.59%
M	118 619.9	6.00%
E	597 343.7	30.21%
S	623 538.5	31.54%

在目前利用条件随机场进行中文分词的研究中，大多数还是处于单一特征标记阶段。对于组合特征也

多是简单特征增加，在实验中发现，不同的特征组合带来的效果是不一样的，并不是特征越多训练出来的模型分词效果就越好，可能出现特征干扰和由于特征过多而带来的冗余信息，导致分词效果下降的现象[17]。

除字位特征以外，针对现代汉语文本常用的其他特征如"字音特征"、"词长特征"，笔者进行了单独实验和组合实验。首先将训练语料选择其中一些特征组合配上合适的特征标记处理为 CRF++能识别的训练语料格式，如表 6 所示。另外，在处理训练语料时候，将要输出的特征放在最后一列，由于本文研究的是分词任务，所以将位置特征放在最后一列。

表 6　添加多个特征之后的训练语料

食品安全语料	字音特征	词长特征	位置特征
媒	mei	2	B
体	ti	2	E
调	diao	2	B
查	cha	2	E
街	jie	2	B
头	tou	2	E
凉	liang	3	B
拌	ban	3	M
菜	cai	3	E
原	yuan	2	B
料	liao	2	E
部	bu	2	B
分	fen	2	E
为	wei	1	S
人	ren	2	B
造	zao	2	E
或	huo	1	S
含	han	1	S
添	tian	3	B
加	jia	3	M
剂	ji	3	E

在训练学习和测试输出阶段，将语料均分为 10 份，然后按 7:3 的比例进行训练和测试，并使用 10 折交叉验证对模型的稳定性进行评估。

对不同特征和特征组合进行训练和测试输出，并对结果进行测评，结果如表 7 所示。从实验结果可以看出，在 4Tag、5Tag 和 6Tag 类型的特征中，4Tag 和 5Tag 的类型的特征分词之后，P、R、F 值普遍高于 6Tag 类型的特征。在对不同特征选择进行训练时，为了确保区分度，均使用如表 8 所示特征模板进行训练，多特征则在特征模板中另起一行同样使用表 8 中的特征模板。

表 7　不同特征组合的分词测评结果

特征选择	P 值	R 值	F 值
4Tag	92.85%	92.89%	92.87%
4Tag+词长	92.74%	92.78%	92.76%
4Tag+字音	92.53%	92.57%	92.55%
4Tag+词长+字音	92.67%	92.69%	92.68%
5Tag	92.85%	92.90%	92.88%
5Tag+词长	92.64%	92.69%	92.67%
5Tag+字音	92.32%	92.38	92.35%
5Tag+词长+字音	92.02%	92.08%	92.05%
6Tag	92.20%	92.11%	92.16%
6Tag+词长	92.09%	92.00%	92.04%
6Tag+字音	92.00%	91.90%	91.95%
6Tag+词长+字音	91.71%	91.60%	91.65%

表 8　基本的特征模板

特征	特征模板	特征描述
C_{-2}	U01:%x[–2, 0]	当前字的前驱第二个字
C_{-1}	U02:%x[–1, 0]	当前字的前驱第一个字
C_0	U03:%x[0, 0]	当前字
C_1	U04:%x[1, 0]	当前字的后驱第一个字
C_2	U05:%x[2, 0]	当前字的后驱第二个字
$C_{-1}C_0$	U06:%x[–1, 0]/%x[0, 0]	前一个字到当前字的转移概率
C_0C_1	U07:%x[0, 0]/%x[1, 0]	当前字到后一个字的转移概率
$C_{-1}C_1$	U08:%x[–1, 0]/%x[1, 0]	前一个字到后一个字的转移概率

(2) 特征模板的构建和优化

CRF++中的特征模板主要用来定义从训练集中提取特征的方法，使用特征模板从训练集中提取到的特征字符串，在 CRF++中，这些特征都是二值函数，函数的输出用来判断这个标签是否要输出"output"中的特征标签。

在特征模板文件中，主要使用的是 Unigram Template，此特征模板第一个字符是 U，每一行(如 U01:%x[–2, 0])代表一个特征，而宏"%x[行位置，列位置]"则代表相对于当前指向的 token 的行偏移和列的绝对位置，如表 8 所示。每一行"%x[行位置，列位置]"生成一个 CRFs 中的点(state)函数：$f(s, o)$，其中 s 为 t

时刻的标签(output), o 为 t 时刻的上下文。使用表 2 所示的特征模板，以表 6 中的语料来说明:

func1 = if (output = B and feature="U040:媒") return 1 else return 0

它是由 U03:%x[0, 0]在输入文件的第一行生成的点函数。将输入文件的第一行“代入”到函数中，函数返回 1，同时，如果输入文件的某一行在第 1 列也是“媒”，并且它的 output(最后一列)同样也为 B，那么这个函数在这一行也返回 1。

在 template 文件中，每一种特征对应一个特征模板(template 中用换行来区分不同的特征模板)，不同的特征模板和不同特征之间的配合使用同样也会影响到分词的效果[18]。本文利用特征选择中效果比较好的 5Tag 特征标记对不同的特征模板进行实验，结果如表 9 所示，可以看出不同的特征模板构建方式对分词效果的影响较大，其中移除二元特征($C_{-1}C_0, C_0C_1, C_{-1}C_1$)，F 值有明显下降，而一元特征的增加和移除对分词效果影响不明显，当增加的二元特征不包含 C_0 (%x[0, 0])时，对分词结果影响不大。

表 9　应用不同特征模板的分词结果

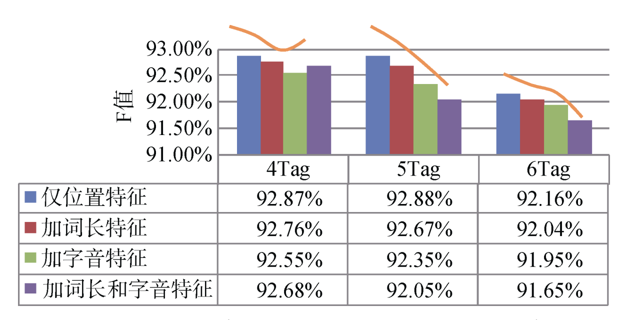
特征模板(对比表 8)	F 值
原始特征模板	92.88%
移除一元特征 C_{-2}、C_2、C_{-1}、C_1	92.72%
移除二元特征 $C_{-1}C_0$、C_0C_1、$C_{-1}C_1$	86.33%
增加一元特征 C_{-3}、C_3	92.73%
增加二元特征 C_1C_2、$C_{-1}C_{-2}$	92.56%

(注：增加的一元特征：U09:%x[–3, 0]和 U010:%x[3, 0]；增加的二元特征：U09:%x[1, 0]/%x[2, 0]和 U10:%x[–1, 0]/%x[–2, 0]。)

4　实验结果分析

对不同特征选择实验结果进行对比分析，由表 7 数据生成图 5，将数据分为三组(4Tag 组、5Tag 组、6Tag 组)进行分析，可以看出，原始的位置特征标记(仅加上位置特征)所得到的分词效果(F 值)最好，加上其他一个或多个特征之后，均有下降趋势。

条件随机场的最大优势是不仅可以融入当前字的各种特征知识，而且可以结合当前字左右特征知识从而形成最有效的特征模板[18]。对不同特征模板的实验结果中，移除当前字和前驱第一个字以及当前字和后驱第一个字的特征行时分词结果变化明显，F 值降低较多。表 10 中，统计了语料中词长分布情况，其中二字词和三字词占 46.62%的比重，所以不难看出当前字和前驱第一个字以及当前字和后驱第一个字的特征行在本文研究的语料自动分词的特征模板构建中是不可缺少的。而由于三字以上词所占比重较少，只有约 2.28%，因此当涉及到前驱第二个字和后驱第二个字，以及前驱第三个字和后驱第三个字的特征模板的变化时，则对分词结果影响不大。



	4Tag	5Tag	6Tag
仅位置特征	92.87%	92.88%	92.16%
加词长特征	92.76%	92.67%	92.04%
加字音特征	92.55%	92.35%	91.95%
加词长和字音特征	92.68%	92.05%	91.65%

图 5　不同特征组合所得 F 值变化趋势

表 10　食品安全突发事件语料中词长分布

词类型	词长度	所占百分比
单字词	1 039 205	51.10%
二字词	841 690	41.39%
三字词	106 307	5.23%
四字词	28 220	1.39%
五字词	8 893	0.44%
六字词	2 626	0.13%
其他	6 598	0.32%

本文针对食品安全突发事件语料的自动分词效果相比于在一些标准测试集上的测试效果略低[19]。分析原因主要在于处理语料时，首先进行机器分词然后人工校正，在后期对错误分词分析的过程中发现，人工校正的过程中出现较多的错误，有许多机器分词错误未予以纠正，也有些机器分词正确而后期人工校正的过程中修改为错误的情况，这对训练学习阶段和模型测评阶段均有影响。

5　结　语

将条件随机场模型应用到食品安全突发事件语料的自动分词中，使用较为成熟、稳定性较强的 CRF++工具对其进行逐一实验，并且考虑到文本的多特征性以及多特征相互组合的可能。实验结果表明特征标记的选择以及不同特征组合的选择会影响到分词效果，其中仅加上位置特征的特征选择 4Tag 和 5Tag 的分词

效果较好, 其 F 值达到 92.87%和 92.88%, 而加上其他特征之后 F 值均有下降。同时, 通过对不同特征模板分词效果的对比分析, 选择符合所选特征和合适本文研究对象的特征模板。

在未来的研究中, 将在文本特征上做进一步挖掘, 找到能将上下文的语义和文本结构信息融合进去的特征, 期望在自动分词上得到更好的效果。

参考文献:

[1] 李洪峰. 食品安全社会共治的现实困境与发展对策[J]. 食品与机械, 2016, 32(4): 234-236. (Li Hongfeng. Analysis of Realistic Plights and Countermeasures in Social Co-governance on Food Safety in China[J]. Food & Machinery, 2016, 32(4): 234-236.)

[2] 王辉霞. 公众参与食品安全治理法治探析[J]. 商业研究, 2012(4): 170-177. (Wang Huixia. Public Participation in Food Safety Management of the Rule of Law [J]. Commercial Research, 2012(4): 170-177.)

[3] 奉国和, 郑伟. 国内中文自动分词技术研究综述[J]. 图书情报工作, 2011, 55(2): 41-45. (Feng Guohe, Zheng Wei. Review of Chinese Automatic Word Segmentation [J]. Library and Information Service, 2011, 55(2): 41-45.)

[4] 张星联, 唐晓纯. 我国食品安全预警数据库系统的建设与实现[J]. 食品科技, 2008, 33(12): 250-254. (Zhang Xinglian, Tang Xiaochun. Establishment on Database System of Food Safety Early-warning in China [J]. Food Science and Technology, 2008, 33(12): 250-254.)

[5] 吴云红, 朱亮, 初炜, 等. 食品监管改革的关键——基于互联网的动态第三方数据库[J]. 食品工业科技, 2009(9): 272-274. (Wu Yunhong, Zhu Liang, Chu Wei, et al. Key of Food Supervision and Administration Reform-dynamic and Third Party Database Based on Internet [J]. Science and Technology of Food Industry, 2009 (9): 272-274.)

[6] 余清, 洪源. 加工食品风险数据库的构建思路[J]. 价值工程, 2013(30): 174-175. (Yu Qing, Hong Yuan. Construction Idea for Risk Database of Processed Food [J]. Value Engineering, 2013(30): 174-175.)

[7] 贾凯, 彭培好, 阮伟玲. 四川省彭州市三界镇农民专业合作社调查研究[J].北京农业, 2014(3): 247-248. (Jia Kai, Peng Peihao, Ruan Weiling. Study on the Investigation of Farmer Cooperatives in Sanjie Town, Pengzhou City, Sichuan Province [J]. Beijing Agriculture, 2014(3): 247-248.)

[8] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3): 8-19. (Huang Changning, Zhao Hai. Chinese Word Segmentation: A Decade Review [J]. Journal of Chinese Information Processing, 2007, 21(3): 8-19.)

[9] Zeng D, Wei D, Chau M, et al. Domain-specific Chinese Word Segmentation Using Suffix Tree and Mutual Information [J]. Information Systems Frontiers, 2011, 13(1): 115-125.

[10] 刘泽文, 丁冬, 李春文. 基于条件随机场的中文短文本分词方法[J]. 清华大学学报:自然科学版, 2015, 55(8): 16-20. (Liu Zewen, Ding Dong, Li Chunwen. Chinese Word Segmentation Method for Short Chinese Text Based on Conditional Random Fields [J]. Journal of Tsinghua University:Science and Technology, 2015, 55(8): 16-20.)

[11] Lafferty J D, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]//Proceedings of the 18th International Conference on Machine Learning. 2001: 282-289.

[12] Pearl J. Bayes and Markov Networks:A Comparison of Two Graphical Representations of Probabilistic Knowledge [R]. Los Angeles, California, USA: University of California, 1986.

[13] Wallach H M.Conditional Random Fields: An Introduction [EB/OL]. (2004-02-24). http://www.inference.phy.cam.ac.uk/hmw26/papers/crf_intro.pdf.

[14] CRF++: Yet Another CRF Toolkit [EB/OL]. [2014-08-04]. http://crfpp.sourceforge.net/.

[15] 中国科学院计算技术研究所. ICTCLAS 汉语分词系统 [CP/OL]. (2016-02-17). [2016-06-30]. http://ictclas.nlpir.org/. (Institute of Computing Technology of the Chinese Academy of Sciences. ICTCLAS Chinese Word Segmentation System [CP/OL]. (2016-02-17). [2016-06-30]. http://ictclas.nlpir.org/.)

[16] 岳金媛, 徐金安, 张玉洁. 面向专利文献的汉语分词技术研究[J]. 北京大学学报: 自然科学版, 2013, 49(1): 159-164 . (Yue Jinyuan, Xu Jin'an, Zhang Yujie. Chinese Word Segmentation for Patent Documents [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 159-164.)

[17] Chen L, Li M, Zhang J, et al. A Double-Layer Word Segmentation Combined with Local Ambiguity Word Grid and CRF [J]. Transactions on Computer Science & Technology, 2013, 2(1): 1-8.

[18] 黄水清, 王东波, 何琳. 以《汉学引得丛刊》为领域词表的先秦典籍自动分词探讨[J]. 图书情报工作, 2015, 59(11): 127-133. (Huang Shuiqing, Wang Dongbo, He Lin. Exploring of Word Segmentation for Fore-Qin Literature Based on the

Domain Glossary of Sinological Index Series [J]. Library and Information Service, 2015, 59(11): 127-133.)
[19] Zhao H, Huang C N, Li M, et al. An Improved Chinese Word Segmentation System with Conditional Random Field [C]// Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing.2006: 162-165.

作者贡献声明:

王东波: 提出研究思路, 设计研究方案;
王东波, 张越, 朱丹浩: 采集、清洗和分析数据;
张越: 进行实验, 起草论文;
王东波, 张越: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1-3]由作者自存储, E-mail: db.wang@njau.edu.cn; 支撑数据[4]见期刊网络版 http://www.infotech.ac.cn。
[1] 张越, 王东波. data.txt. 食品安全突发事件汉语分词训练和测试数据.
[2] 张越, 王东波. Template. 食品安全突发事件汉语分词特征模板.
[3] 张越, 王东波. result.txt. 食品安全突发事件汉语分词结果.
[4] 张越, 王东波. wordseg 的 java 工程文件.

收稿日期: 2016-09-22
收修改稿日期: 2016-10-31

Segmenting Chinese Words from Food Safety Emergencies

Zhang Yue[1] Wang Dongbo[1, 2] Zhu Danhao[3]
[1](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)
[2](Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095, China)
[3](Library of Jiangsu Police Institute, Nanjing 210031, China)

Abstract: [**Objective**] This paper examines the automatic word segmentation models, which plays key roles to build databases for food safety administration. We used the statistical learning method based on conditional random field to segment words from food safety emergencies. [**Methods**] First, we analyzed the length of target words and conducted multiple experiments on the selection and template of word features for the automatic segmentation methods. Second, we identified the impacts of different features and templates to the segmentation results. [**Results**] We found that selecting more features might not yield better results due to the characteristics interference. About 46.62% of the phrases from the corpus of food safety emergencies only contained two or three words. The first words before and after the current word of the features template pose more effects to the results. [**Conclusions**] We have identified the optimal feature and template for the automatic segmentation of words and the F score reaches 92.88% with the 5Tag features.

Keywords: Chinese Word Segmentation　Food Safety　Conditional Random Field　Feature Template　Feature Selection